Event Builder Review

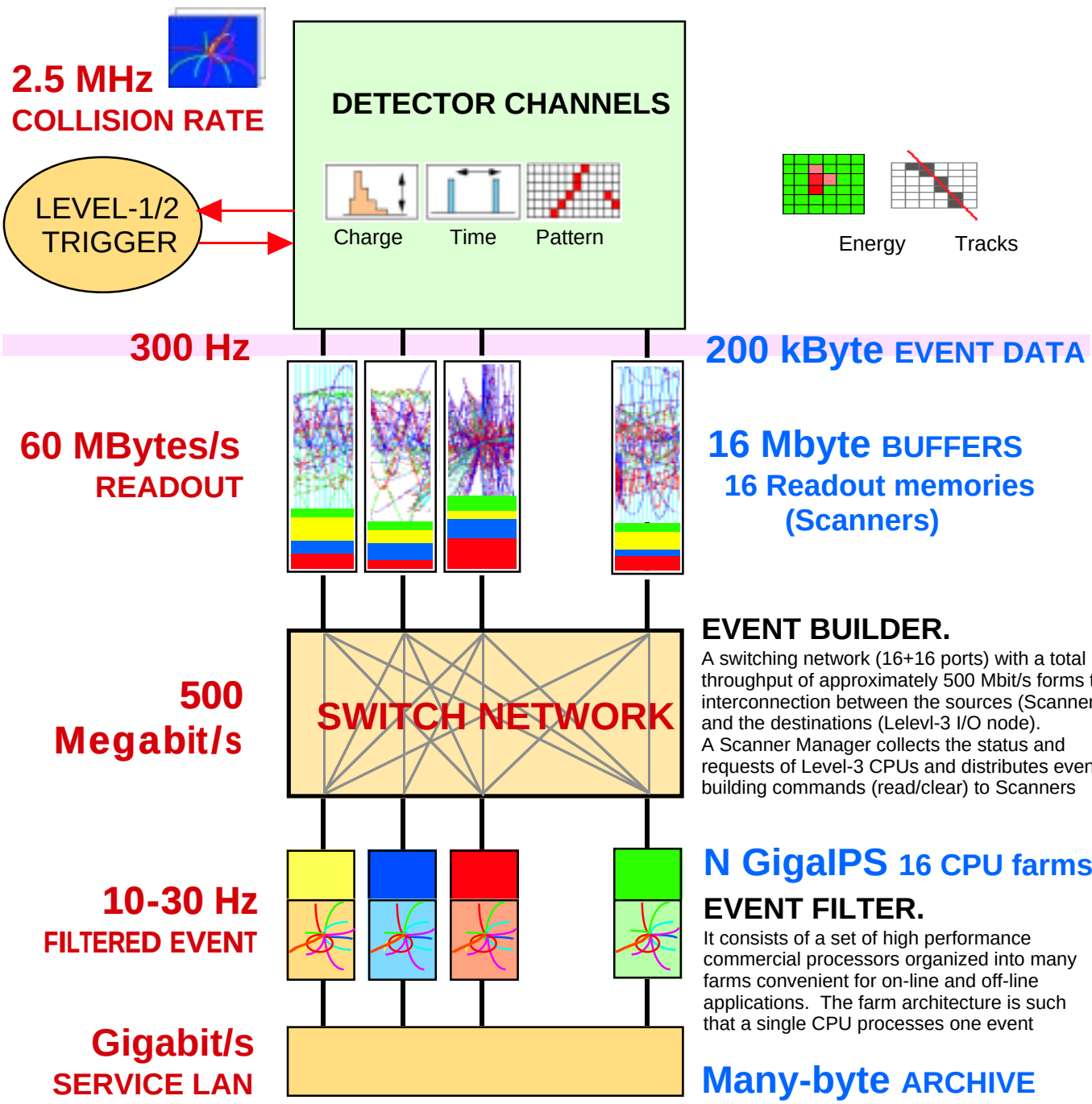
Outline

- Introduction
- Run I System
- Switches
- Event Building
- The ATM/FC Stands/Prototypes
- Is ATM the right technology?

Introduction

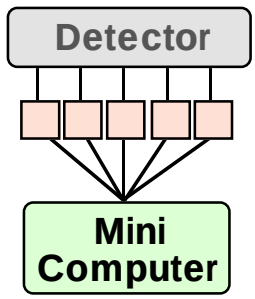
- DAQ: the goal
- DAQ evolution
- Trigger/DAQ evolution
- Technology evolution

DAQ: the goal

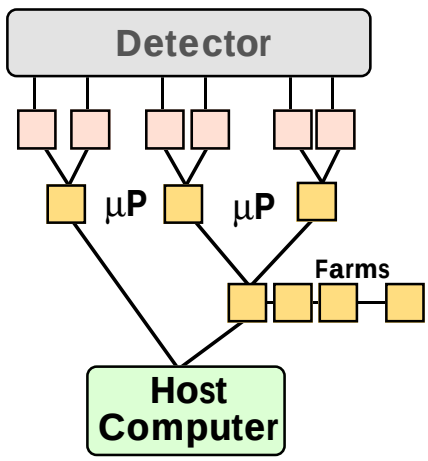

2.5 MHz
COLLISION RATE
DETECTOR CHANNELS
LEVEL-1/2
TRIGGER
Charge
Time
Pattern
Energy
Tracks
300 Hz
200 kByte EVENT DATA
60 MBytes/s
READOUT
16 Mbyte BUFFERS
16 Readout memories
(Scanners)
SWITCH NETWORK
500
Megabit/s
EVENT BUILDER.
A switching network (16+16 ports) with a total throughput of approximately 500 Mbit/s forms the interconnection between the sources (Scanners) and the destinations (Lelevl-3 I/O node).
A Scanner Manager collects the status and requests of Level-3 CPUs and distributes event building commands (read/clear) to Scanners
N GigaIPS 16 CPU farms
10-30 Hz
FILTERED EVENT
EVENT FILTER.
It consists of a set of high performance commercial processors organized into many farms convenient for on-line and off-line applications. The farm architecture is such that a single CPU processes one event
Gigabit/s
SERVICE LAN
Many-byte ARCHIVE

DAQ Technology/architectures

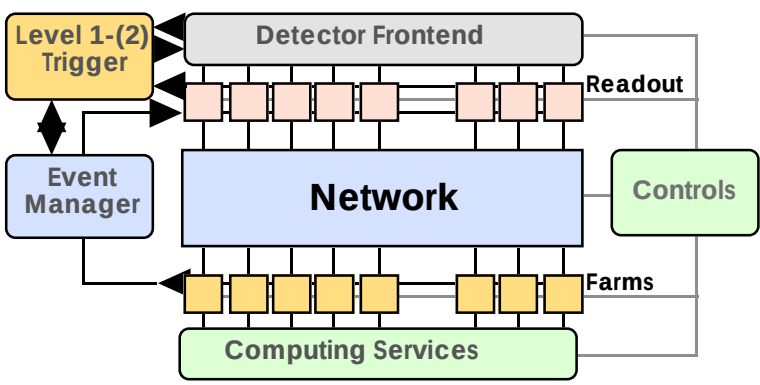
1970-80 MiniComputers
- CAMAC: first standardization
  - kByte/s



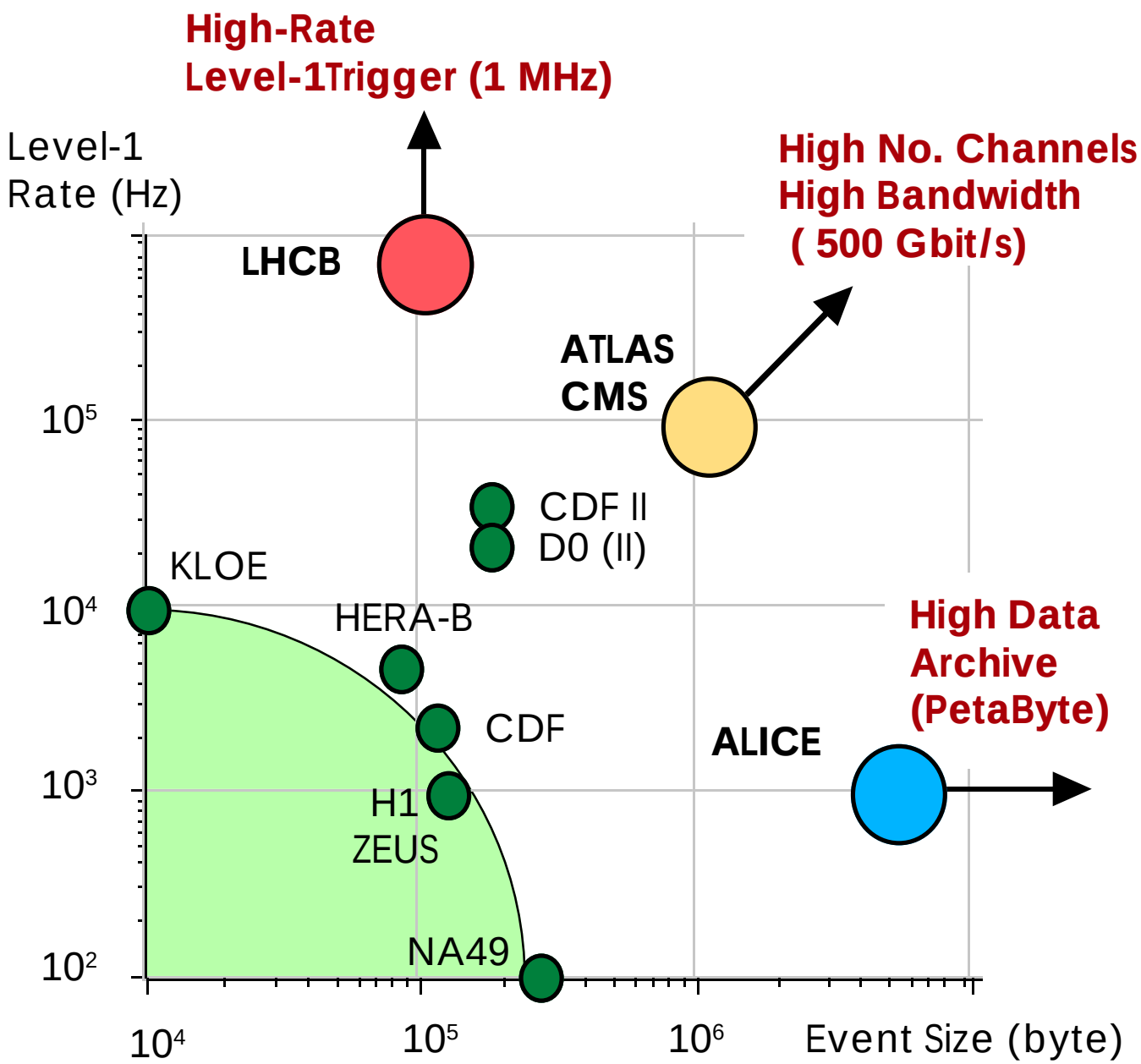
1980-90 MicroProcessors
- Parallel systems
- Distributed intelligence
  - MByte/s



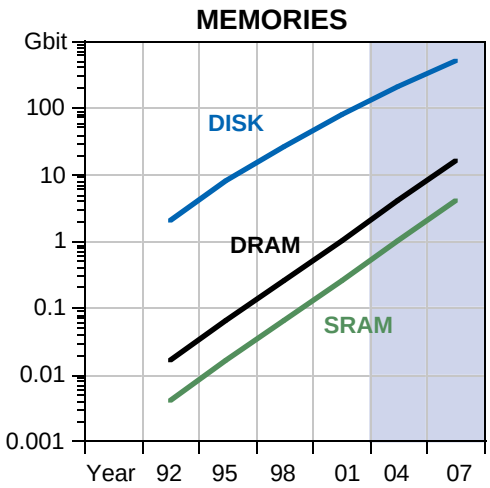
90-2000 Communication
- Embedded processing
- Data and control networks
  - GByte/s

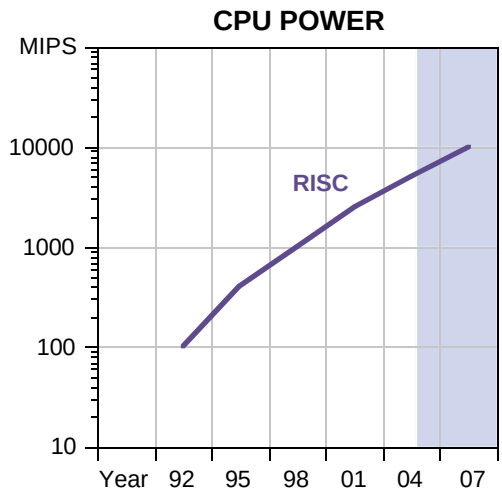


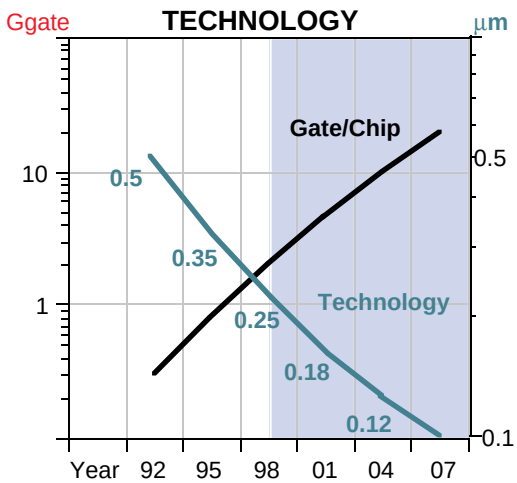
Trigger/DAQ evolution

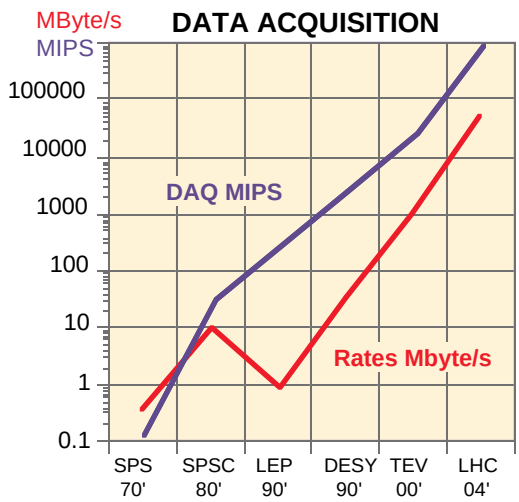


Technology evolution









The CPU processing power increases by a factor 10 every 5 years

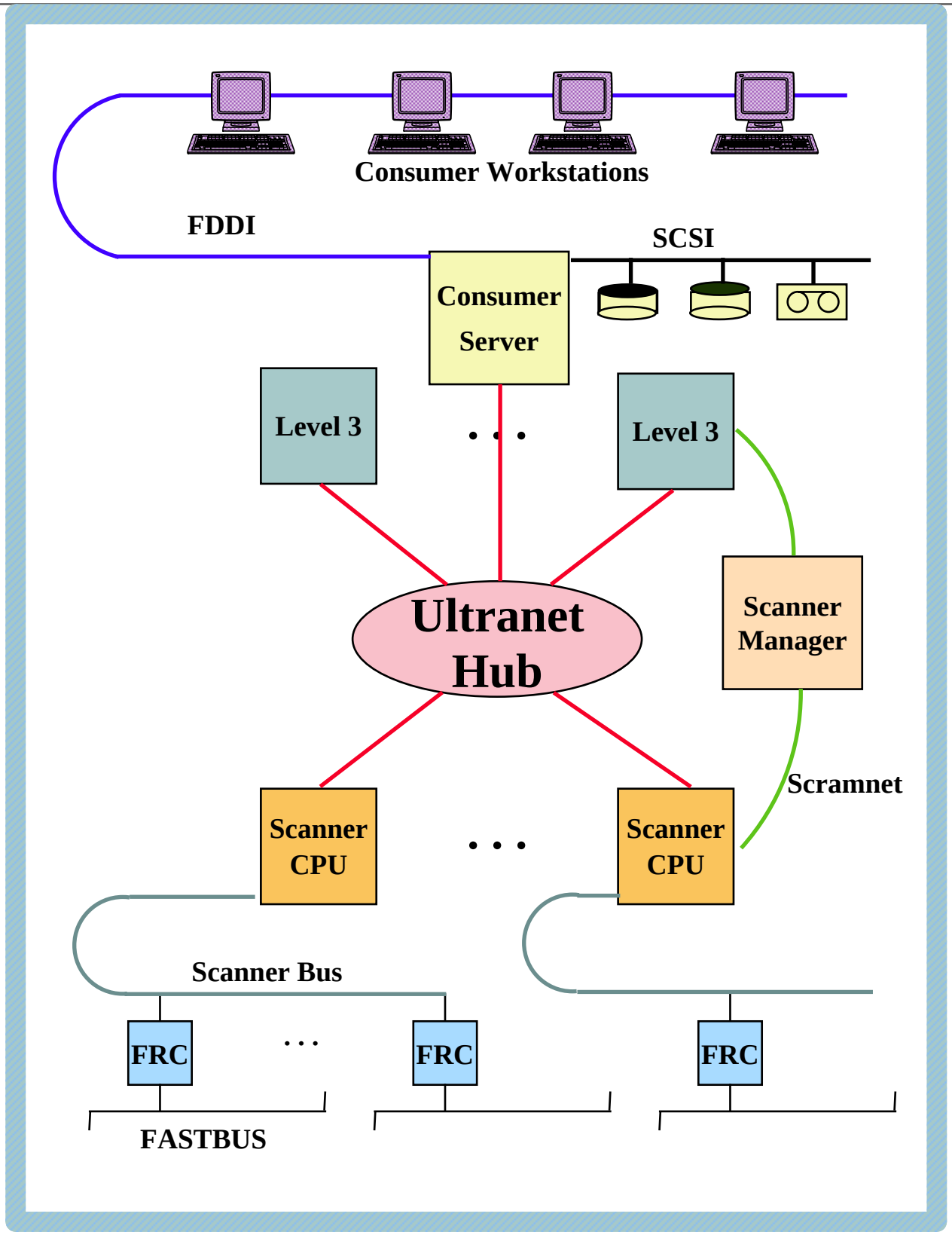
Memory density increases by a factor 4 every two years

The 90's is the data communication decade

Run I System

- Run I Architecture (I)
- Run I Architecture (II)
- Run I Architecture (III)
- Run II Architecture
- Operating mode; Partitioning
- Readout (Scanner) Crate
- Scanner-Switch independence

Run I Architecture (I)



Run I Architecture (II)

- **Gather Fastbus Data into**
  6 VME crates

- **Send Data to Processor Farm in unassembled form:**
  let farm build the event.

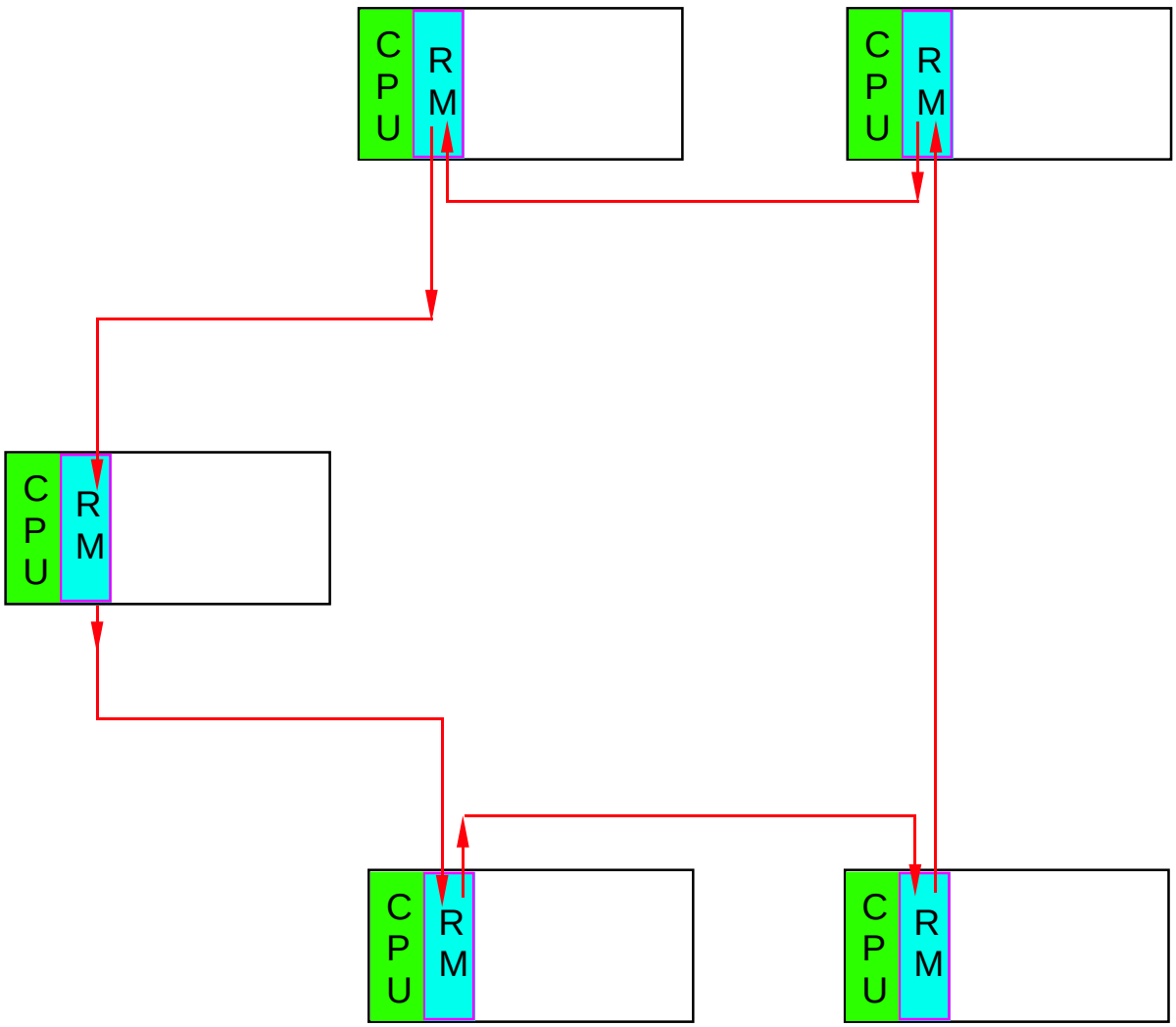
- **Data path between Scanners and SGis:**
  commercial network: ULTRANET

- **Use a centralized intelligence for**
  Event Flow Control

$\Rightarrow$ **Need fast control path,**
use Reflective Memories

Run I Architecture (III)

Special Hardware Detects any write access to the memory and sends write action over optical network
Key parameter: point-to-point drop of 1 μsec



Sample Network of 5 VME Crates, connected via 5 Reflective Memories (RMs).

Run II Architecture

Keep heart of the system:

- **Read out VME modules**
- **Reflective memory**
- **Centralized intelligence for control:**
  Scanner Manager

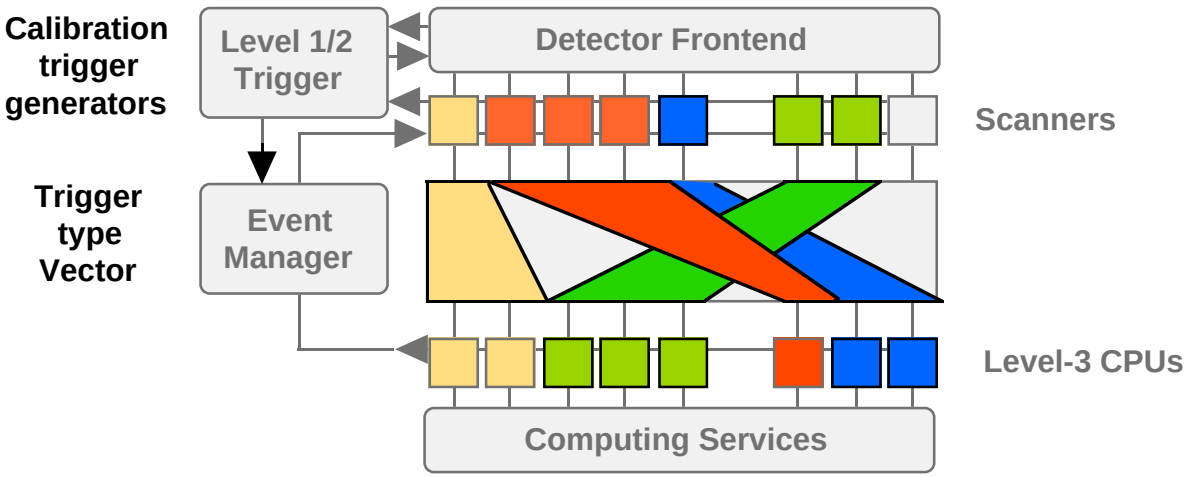
Replace:

- **switch (ULTRANET)**
- **(some) software**
  (including communication with TSI, etc)

(Obvious) things to keep in mind:

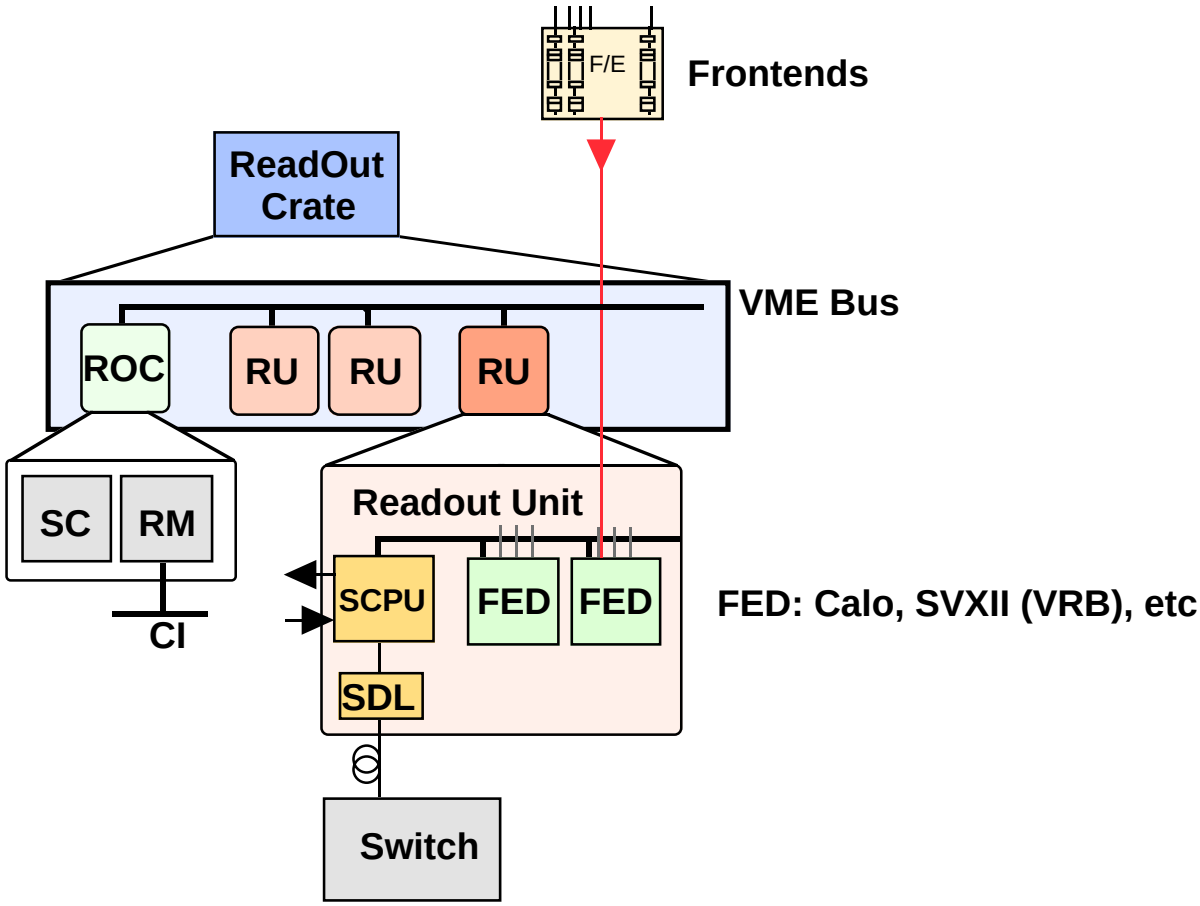
- **Performance**
- **Scalability**
- **Maintainability**

Operating mode; Partitioning



Farm allocation during calibration and test sessions determines (via the event manager) the use of the switch

Readout (Scanner) Crate



- In principle one can have more than one CPU per crate (only makes sense if overheads large compared to DMA capability on VME)
- Actual configuration:
  - One RU (Scanner + FEDs) per crate
  - SCPU acts as system controller (SC) also
  - Reflective Memory (RM) for communication with Manager

Scanner-Switch independence

- Many different switching technologies
- Even after we narrowed it down to 2 (ATM & FC) we needed a modularity that would allow the selection of the switch at 11:55 pm (assuming Run II will start at 00:00)
- Solution: adopt PCI standard, in PMC formfactor

Basic PCI parameters

- Clock speed at 33 MHz; bus width @ 32 bits
- Open-ended (counts on reflection at end)
- Short (because of c)
- Introduced by Intel, now almost everywhere
- PMC: a standard plug-in connector
- Upgradeability: 64 bit out; 66 MHz coming (hardware must change; but software not)

→ Selecting a VME CPU that has a PMC connector enables one to change PMC-to-Switch cards and go on

Selected: Motorola MVME X60Y (X=1,2; Y=3,4)

PS. We also looked at one more manufacturer (Radstone) bottom line: not as good/cheap etc.

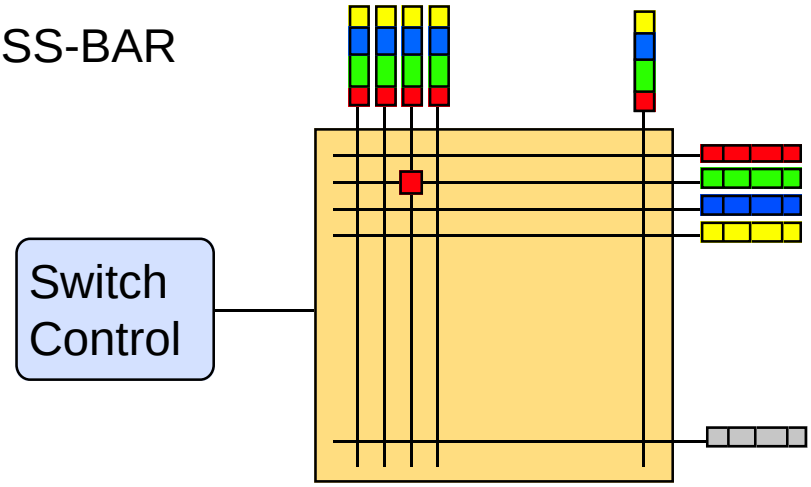
Switches

- Switches: basic types
- Switches: technologies
- Switches: ATM & FC
- Switches: Gbit Ethernet & SCI
- Switches: Others...
- Switch Interface cards
- Switches: Run II "R&D"

Switches: basic types

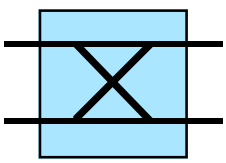
Two basic categories

CROSS-BAR

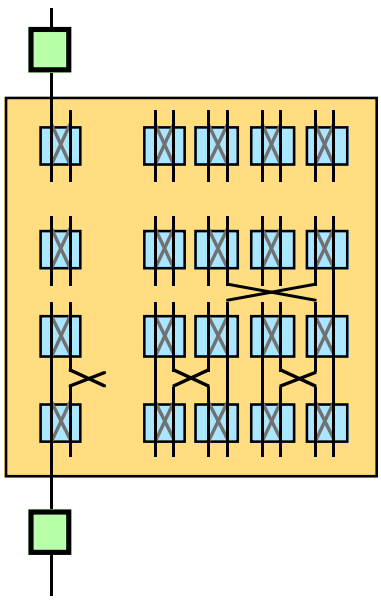


Protocols:
- External control
- Node autorouting
- .....

SWITCH FABRIC



Eg. basic component
2•2 switch unit

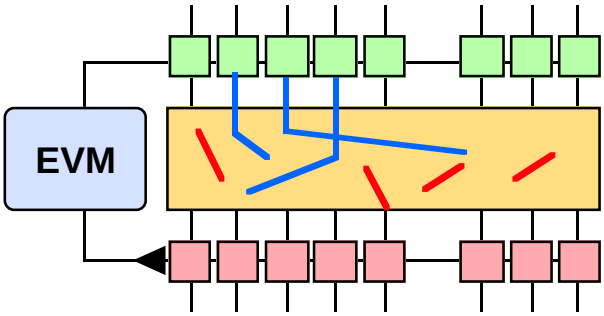


- Packet switching
- Trafic shaping
- Backpressure
- Multipath
-.....

Assumption #1:
we will not develop a CDF-specific switch
→ use commercially available one
→ should have multiple sources also
(remember ULTRANET...)

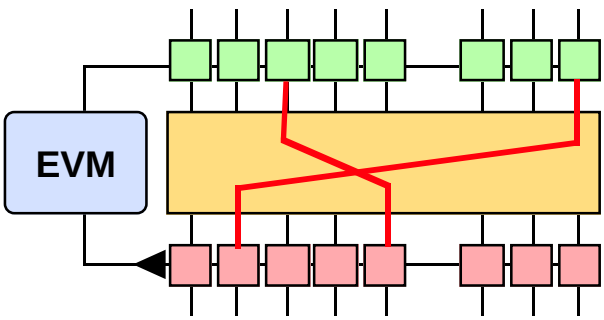
Switches: technologies

- **ATM: Asynchronous Transfer Mode**
  Used by tellecommunications industry



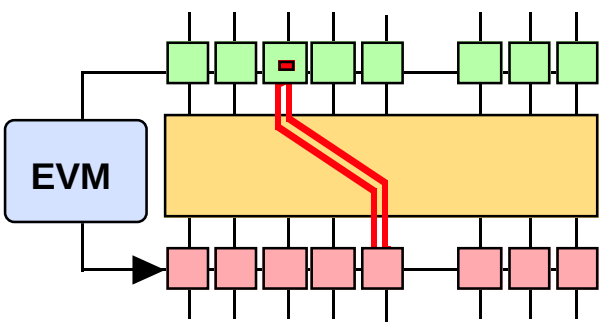
Autorouting (PACKET SWITCHING)
- E.g. ATM and FCS (high classes)
- Any size event fragments
- Adapter layer is complex
- Large system performance ? Latency?
- Data congestion controls ?
(traffic shaping, back pressure.)

- **FibreChannel + GigaEthernet:**
  Used for Computer Connections



Data channel (CIRCUIT SWITCHING)
- E.g. FCS, HIPPI, custom
- Large size fragments (multi-events)
- Cross bar like switches.
Open connection and send data
- Channel auto selection or from central system (e.g. barrel switch)
- Large system cross bar ?

- **SCI:**
  Transparent access to data (looks like cache)



Multi-port memory.
- e.g. MPP architecture, SCI
- No event flow control. Move data only when needed

- Large system feasible? Latency?

Switches: ATM & FC

- **ATM:**

Pros:

- Backed by industries speaking of ~ 100 Mbit/s to homes and 620 Mbit/s in WANs
- Has built-in features (e.g. ABR) which come in handy
- Has been around for ≈ 3 years

Cons:

- Has been around for ≈ 3 years
- Current "standard" (cheap) ATM cards run @ 155 Mbit/s
- ATM Standard speaks of 620 Mbit/s, 1.2 Gbit/s, 2.4 Gbit/s but no 1.2 products yet
- Today: only PCI-ATM card @ 620 Mbit/s from Sun (costs 3,995 $ list price; 155 Mbit/s card: 900 $)

- **FibreChannel:**

Pros:

- Backed by computing industry
- A natural for output from Level-3/storage etc (uniformity)
- Current cards run @ 1 Gbit/s (6 X ATM)

Cons:

- Large overheads (computer connection)
- Switches are (typically) blocking (very few switches with non-blocking architecture for FC classes 2/3)
- FC Standard speaks of 2 and 4 Gbit/s upgrades but no products out yet (may even skip 2Gbit/s)
- Fairly pricy (2,000 for double-buffer cards)

Switches: Gbit Ethernet & SCI

- **Gbit Ethernet:**

Pros:

- Huge existing market
  (90% of the non-modem internet market)
- Stole from Fibrechannel the best features
  (runs @ 1.25 Gbit/s)
- Cards available for $2,000 today
  (and dropping very, very fast)

Cons:

- Has been around for a few months only
  (current switches etc, are manufacturer-specific)
- Doesn't have simultaneous sending of more than one channel
  (like ATM and FC high classes) built-in

- **SCI:**

Pros:

- Transparent access to data
  (every programmer's dream)

Cons:

- Has not captured the market
- Too (?) exotic...

Switches: Others...

- **Optical Switches, etc**

Pros:

- Elegent idea

Cons:

- Market is virtually zero
- Practical problems
  (alignment/frequency modulations etc...)

- **Bus-based switches I**
  (e.g. Sebring ring for PCI)

Pros:

- Huge potential bandwidths (~ 4 GBytes/s)

Cons:

- VERY new (no modules, just chips for now)

- **Bus-based switches II**
  (e.g. Raceway on VME etc)

Pros:

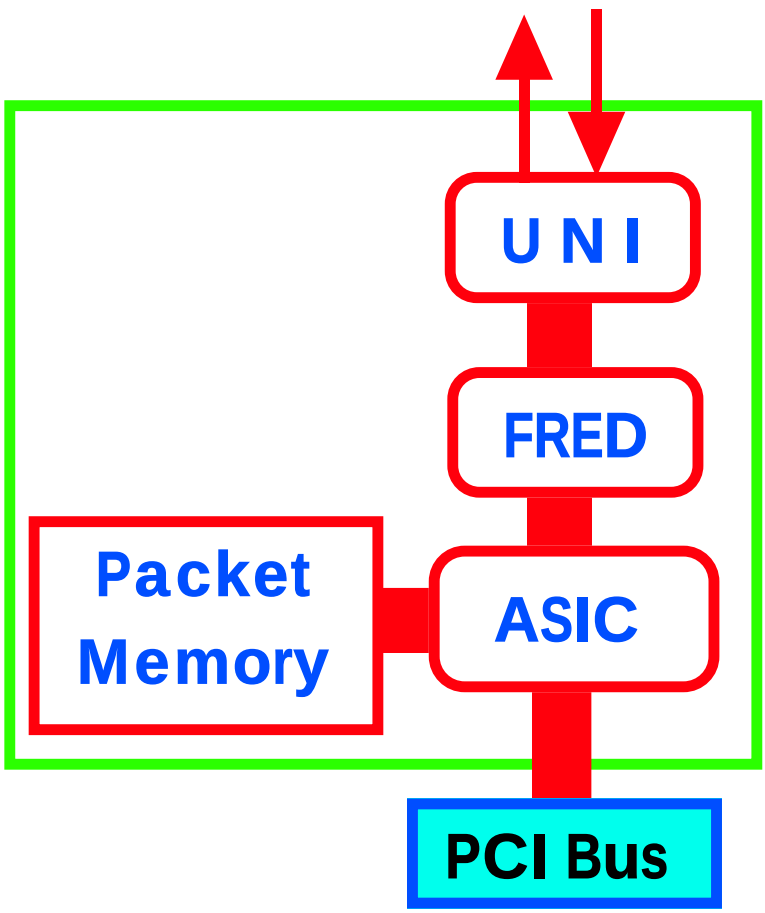
- Very efficient, very scalable

Cons:

- Single source, noone has followed them yet...

Switch Interface cards

- Interphase i4515:
  a PCI to ATM card
  (PMC formfactor)
- 155 Mbit/s



- **To send data:**
  - Load Packet Memory
  - Instruct Board to Send
- **To receive data:**
  - Instruct Board (in advance)
  - Receive Data either in PM or in PCI memory
- **All DMAs by ASIC**
  - PCI to i4515: 46 MB/s
  - i4515 to PCI: 64 MB/s

Switches: Run II "R&D"

VIP #1:

Assumption #1 (reminder):
we will not develop a CDF-specific switch
→ use commercially available one
→ should have multiple sources also
(remember ULTRANET...)

Follow the industry → concentrate on
ATM, Fibrechannel, Gbit Ethernet

VIP #2:

History (reminder):
the current program started in 1994

There was no Gbit Ethernet in 1994
→ ATM, Fibrechannel

VIP #3:

Facts (reminder):
not enough person-power, $ for both

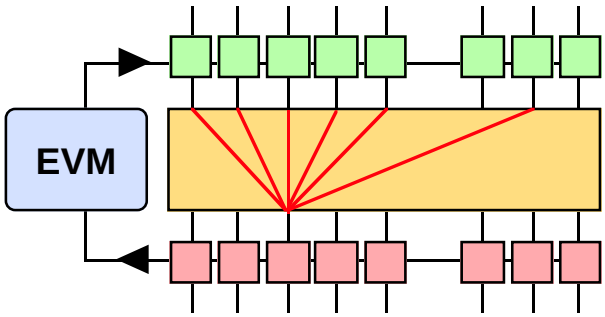
Split work with CERN: they do FC
→ ATM

Event Building

- Event Building: the ultimate bottleneck
- Traffic Shaping: Barrel-Shifter
- Traffic Shaping: Rate Division
- Traffic Shaping: Switch-based
- Event Building: Summary

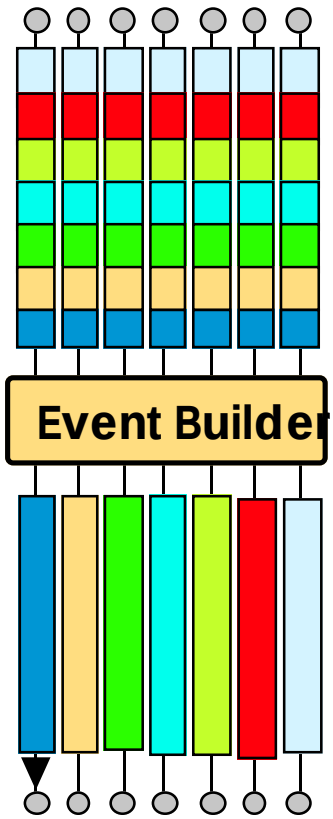
Event Building: the ultimate bottleneck

Event Manager/Supervisor/Allocator:
source → destination mapping



Remaining problem:
ALL sources must communicate with single destination (want the entire event...)

16+16 ports ≈ 500 Mbit/s



Readout units

Event fragments in readout buffers

Full events in CPU memories

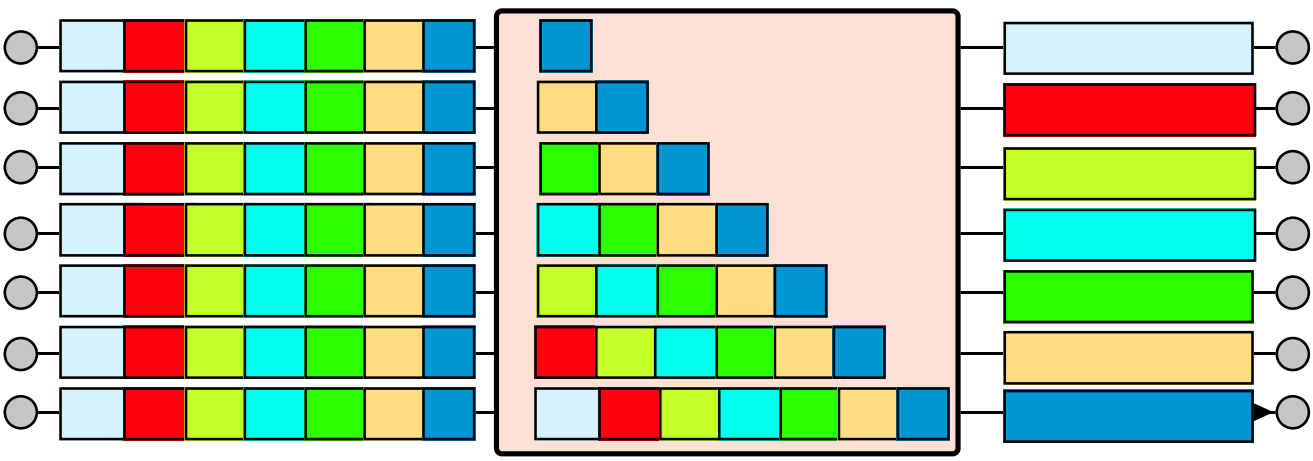
Event filter CPUs

Solution: "traffic shaping"
Four types:
- **(a) barrel shifter**
- **(b) rate division**
- **(c) switch-based**
- **(d) data on demand**

Traffic Shaping: Barrel-Shifter

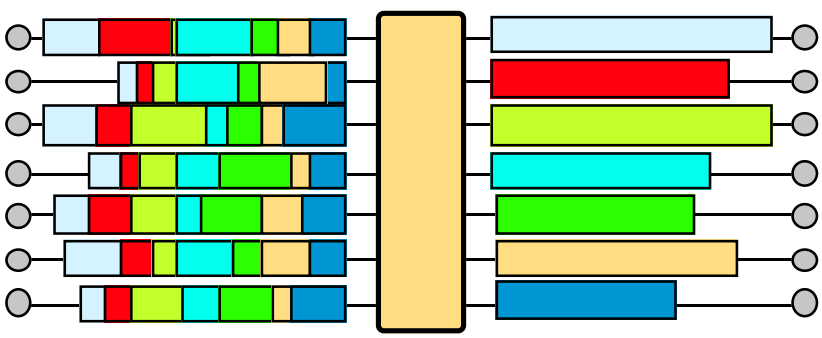
Basic Idea:

Round 1:
Source 1 → Destination 1
...
Round n:
Source 1 → Destination n
Source 2 → Destination n-1
...
Source n → Destination 1



Problems/issues:
- **Control timing of each "slice"**
- **Uneven Event sizes...**

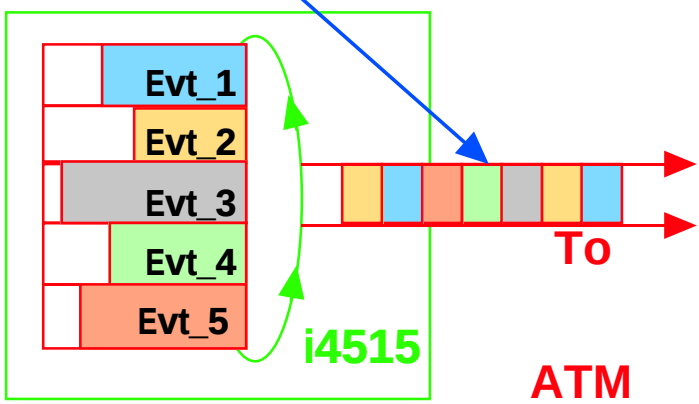
If solved by EVM, then overheads



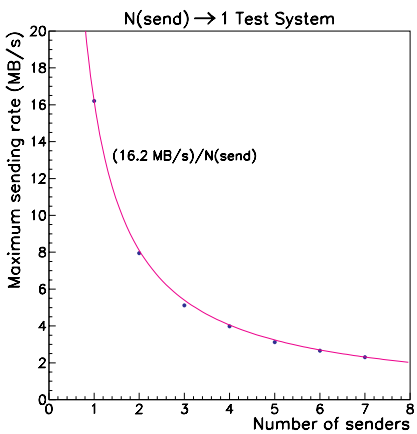
Traffic Shaping: Rate Division

Basic Idea:

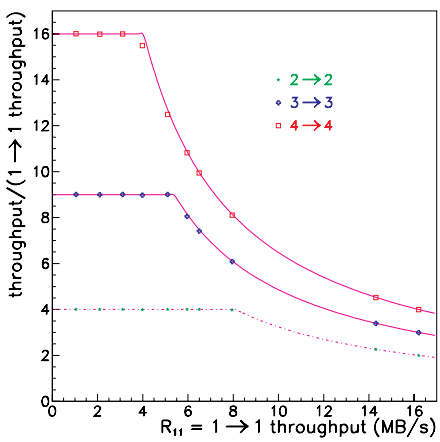
- Simultaneously send to n destinations at 1/n of the speed
- In practice: "simultaneous" means barrel-shifting with small data size (e.g. ATM: 53 bytes)



- **The ultimate barrel-shifter:**
  - (a) Size(basic transfer) δ = 48 bytes
  - (b) Size(event fragment) S = 2000 byte $\rightarrow$ δ/S << 1 $\rightarrow$ very efficient



First results from ATM switch

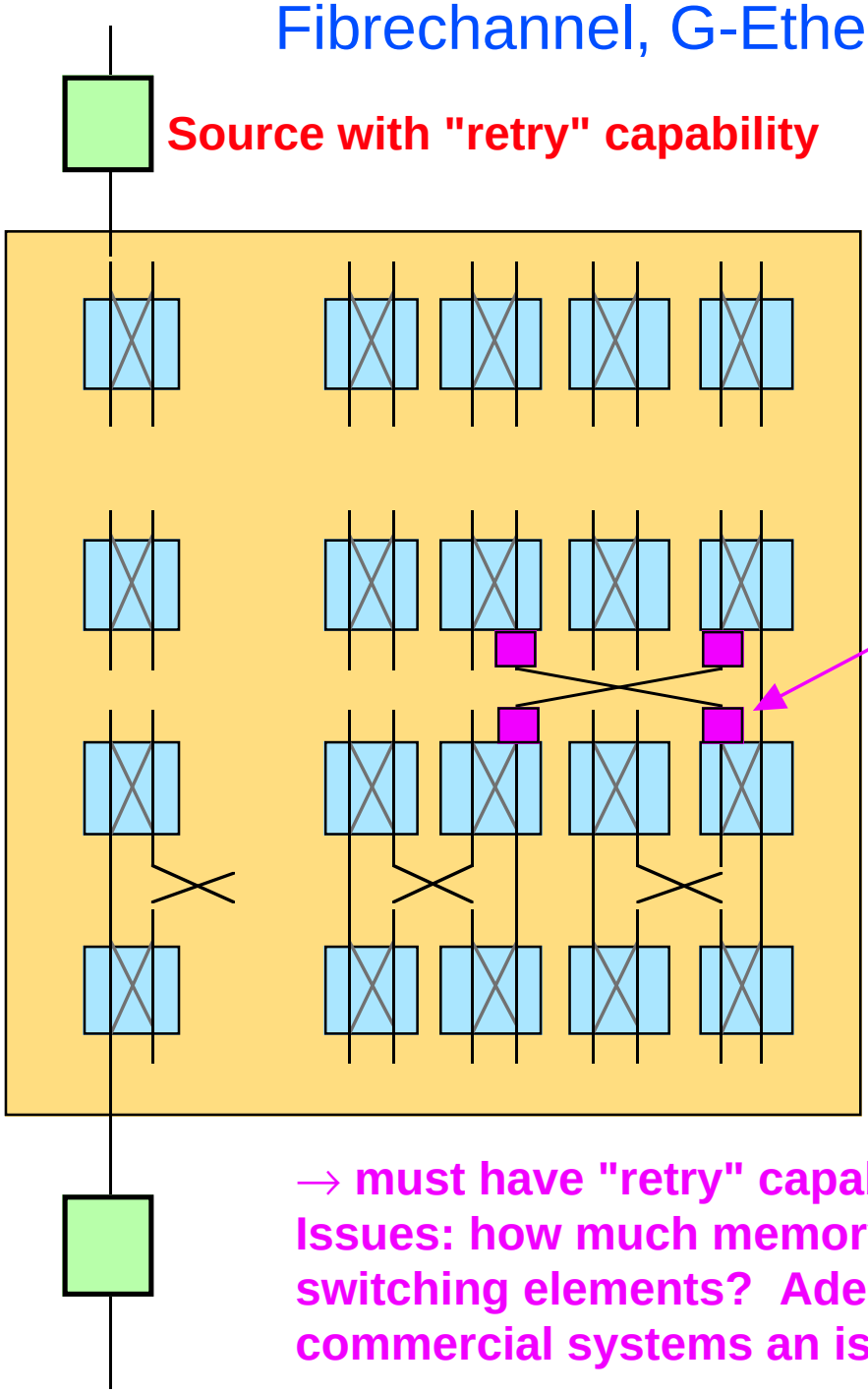


- Points: Max. sending rate for no cell losses to be seen (measured);
- Line: fit to $V = V_{max}/(\varepsilon + N_{send})$
  - ε consistent with 0 (no extraneous congestion)
  - V_{max}=16.2 MB/s

System is *by definition* linear in N_{send}

Traffic Shaping: Switch-based

Fibrechannel, G-Ethernet...



→ must have "retry" capability...
Issues: how much memory in between switching elements? Adequacy of commercial systems an issue...

Final Traffic shaping: ask for data when needed (e.g. SCI, PCI rings, etc...)

Event Building: Summary

- Many technologies with different merits/drawbacks
- Need Event "mapper" or data-defined mapping
- Bottlenecks: different nature depending on switching protocol
- Different from off-the-shelf commercial network router: almost none of current commercial intranets operate at N_{users}*10 Mb/s
- System characterized by
  - Switch technology
  - Input/output modules

The ATM/FC Stands/Prototypes

- FNAL EVB Testbench
- EVB Testbench: Phase I
- EVB Testbench: The ATM switch
- EVB Testbench: Scanner Crate
- EVB Testbench
- Point-to-Point tests
- PtoP tests; ATM
- PtoP tests; FC (I)
- PtoP tests; FC(II)
- FC Switches
- Switch Overhead
- FC Summary
- Linearity/Overheads
- Comparison ATM/FC
- Parenthesis: Gbit Ethernet

FNAL EVB Testbench

Goals:

1. Learn ATM
2. Build full DAQ System based on ATM
3. *If it works, use for CDF Run II*

People:

MIT: T. Daniels, K. Kelley, P. Ngan, P. Sphicas, T. Shah, J. Tseng, S. Tether, D. Vucinic

FNAL: E. Barsotti, M. Bowden, J. Patrick

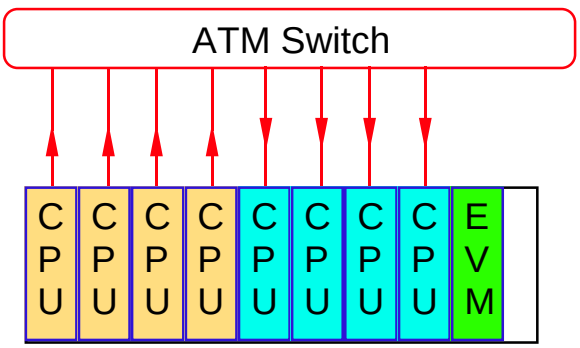
Resources ($) from

— FNAL Computing Division
— CDF Upgrade
— US_CMS

Plan:

— We do ATM
— CERN-CMS does Fibrechannel
— Compare and choose the best

EVB Testbench: Phase I



CPU — PPC VME board: RU

CPU — PPC VME board: FU

EVM — PPC VME board: Evt Manager

1. **Install/operate:**
   Switch, CPUs, ATM interface cards
2. **All CPUs (PowerPC "PPC") in single VME crate**
   Use VME as the "Control network"
   4 CPUs act as Scanners
   4 CPUs act as Level-3 nodes (or I/O nodes)
   1 CPU acts as Event Manager
3. **Port Software from Run I DAQ**
   – No crate readout for Switch Inputs
   – Disable evt processing in L3
4. **Create I/O driver for PCI–ATM cards**
5. **Run 4 × 4 DAQ prototype in single crate**

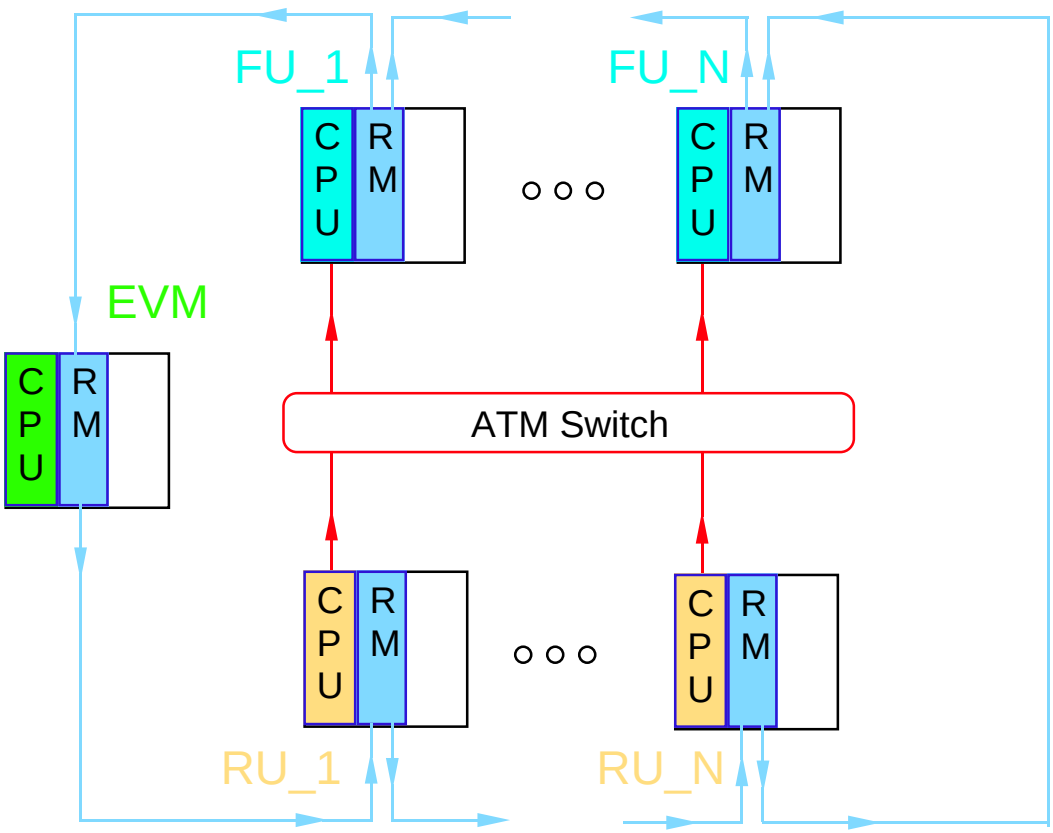
EVB Testbench: The ATM switch

- Forerunner ASX-1000
- Installed in CDF counting room
- Originally equipped with 8 ports (4 inputs + 4 outputs)

- **Port speed: 155 Mbit/s**
- **Can be used for 622 Mbit/s links by multiplexing 4 ports**

EVB Testbench: Scanner Crate

- **Next Step:**
  decoupled CPUs:
  one per crate
- **Control Network:**
  Reflective Memory

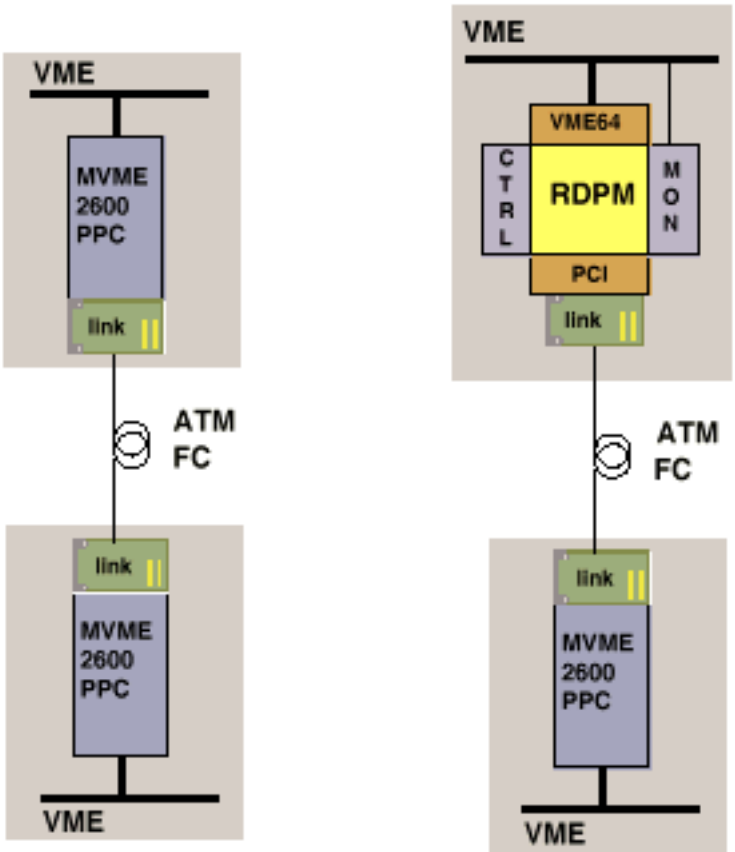


EVB Testbench

Installation in CDF "Level-3" counting room
Similar (but Fibrechannel) installation @ CERN

Point-to-Point tests

Senders: MVME CPU (like CDF) or special (CMS-specific) module

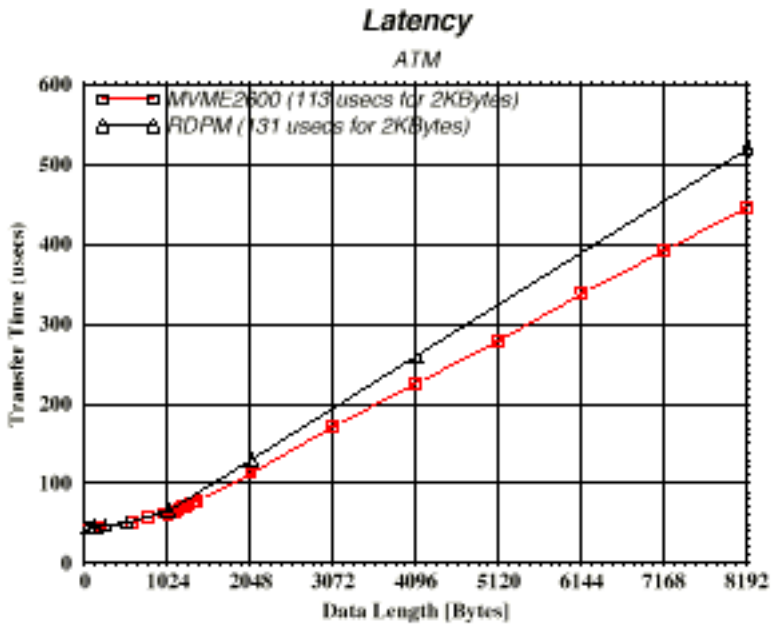


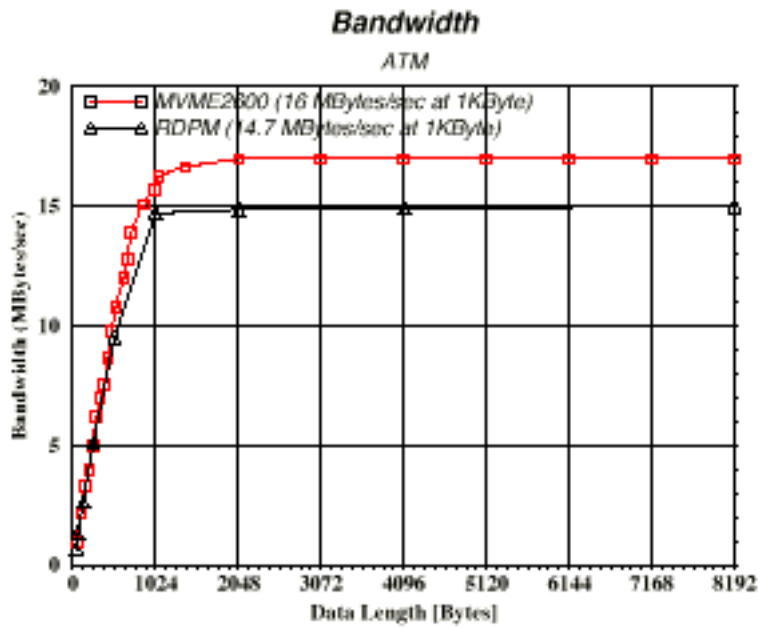
Interface Boards:

- ATM : INTERPHASE (155 MBits/sec)
- FC : SYSTRAN FibreXPress (1062 MBits/sec)
  SYSTRAN FibreXPress GOLD (prototypes, FAST)

Note; later on switch will be ANCOR (first generation). Did not do FC/2-3 in hardware (only software emulation). Current tests only FC Class 1/2

PtoP tests; ATM





Asymptotic values idential to CDF results.

Rise is faster (no operating system).

Recently: installed VxWorks. Results now identical.

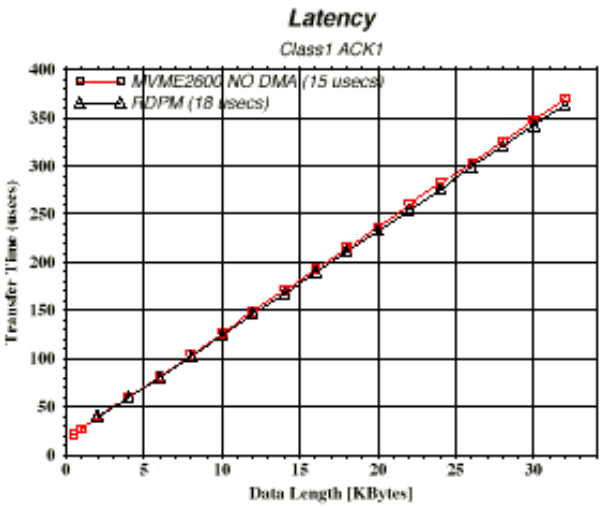
PtoP tests; FC (I)

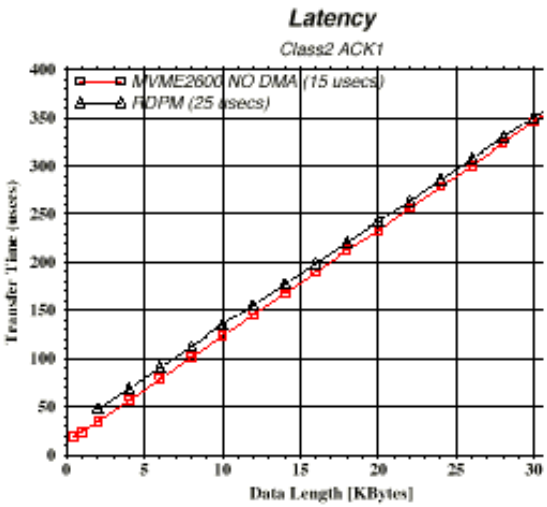
FC protocols:

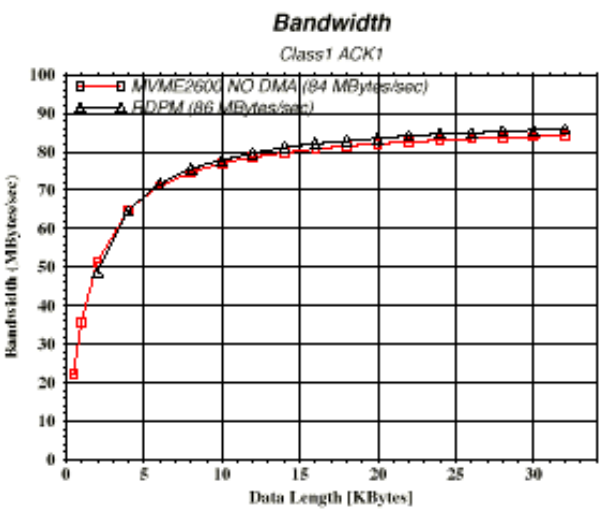
- Class 1: Dedicated connection between communicating ports
  Max. bandwidth available between two ports
  Connection until a disconnect request is received

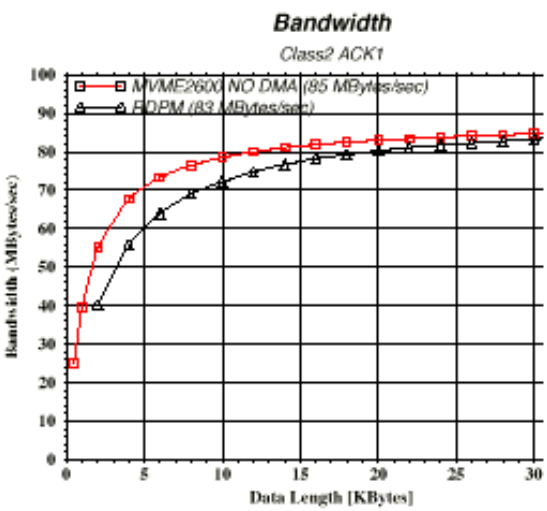
- Class 2: Connectionless service
  Frames are multiplexed at frame boundaries

- Frame Size is 512 Bytes (Switch only supports 512 Bytes frames)

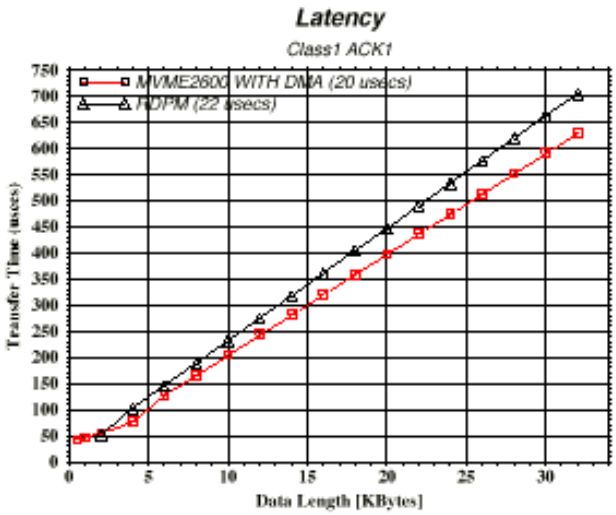


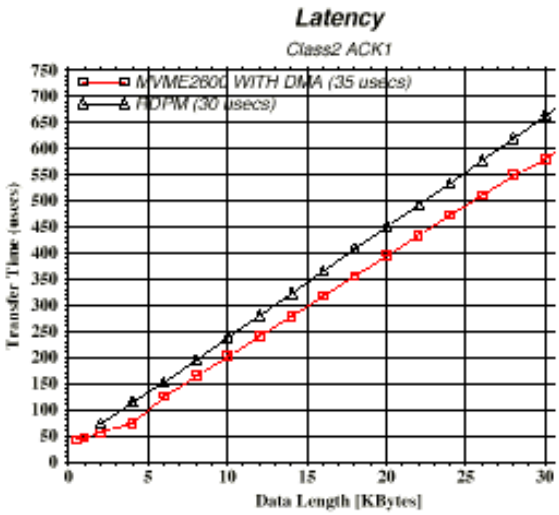


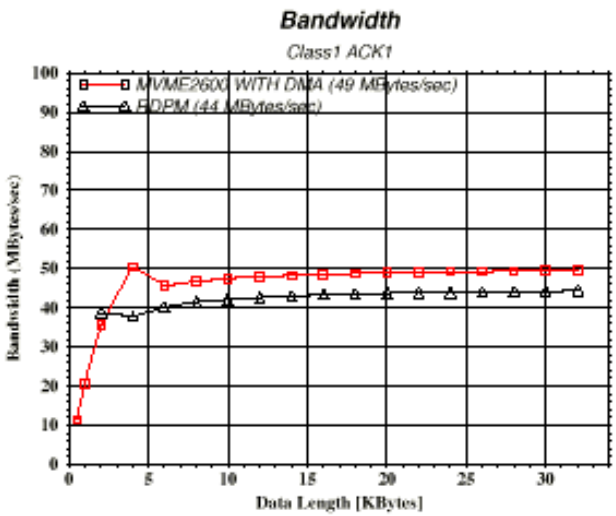


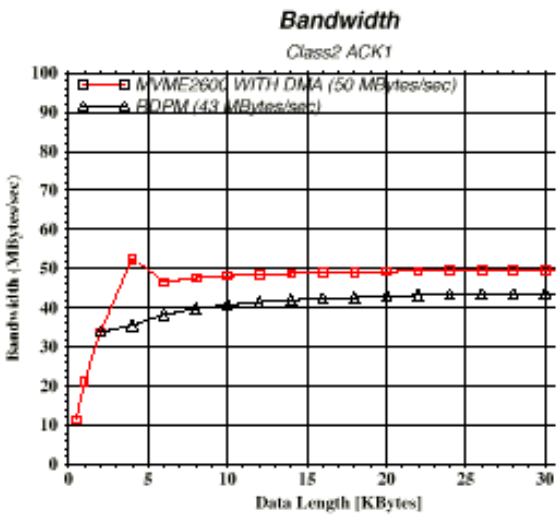


PtoP tests; FC(II)



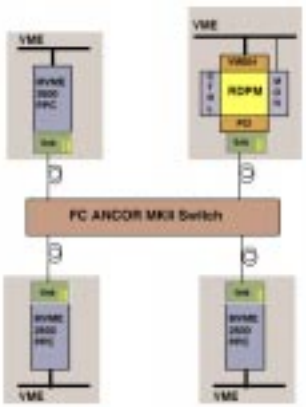






Note: cheap version of cards is single-ported (equivalent to single-buffer, not dual-buffer)
Get only 1/2 of max speed

FC Switches



- **ANCOR FCS 1062 Switch**

  - received in **Nov. 96**
  - Class 1 at full speed
  - Class 2/3 switching done by **software** (1 MByte/sec)
  - frame size = **128 bytes**
  - **compatibility problems** with Tachyon 2.0 chip (optical problems)

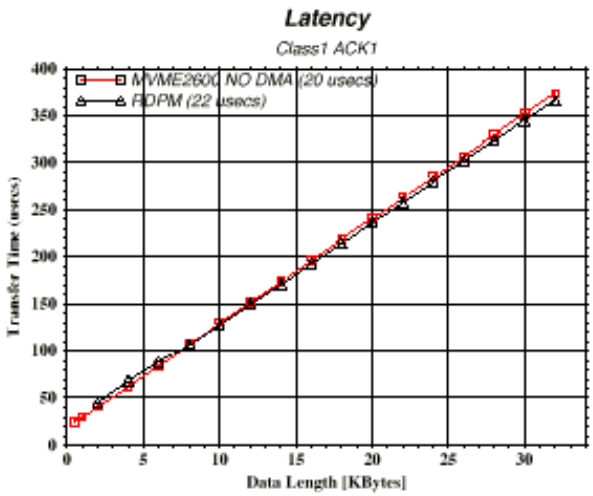
- **Brocade Switch**

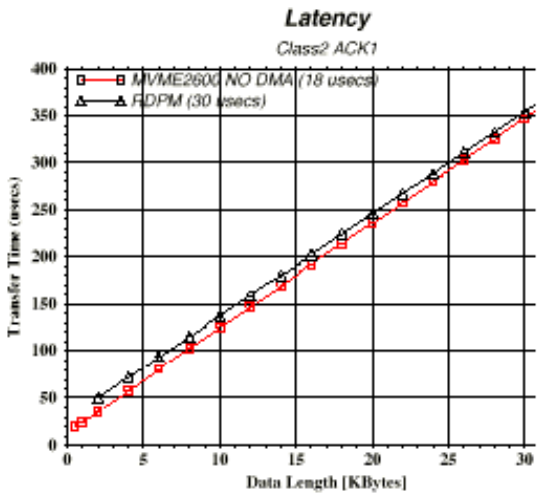
  - on loan for a few day in **Sep. 97** (ATLAS collaboration)
  - NO Class 1
  - Class 2/3 at full speed
  - frame size = **2048 bytes**
  - **NO** compatibility problems with Tachyon 2.0 chip

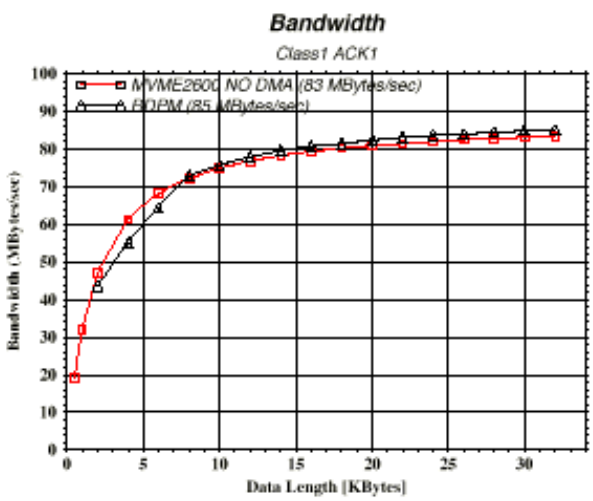
- **ANCOR MKII Switch**

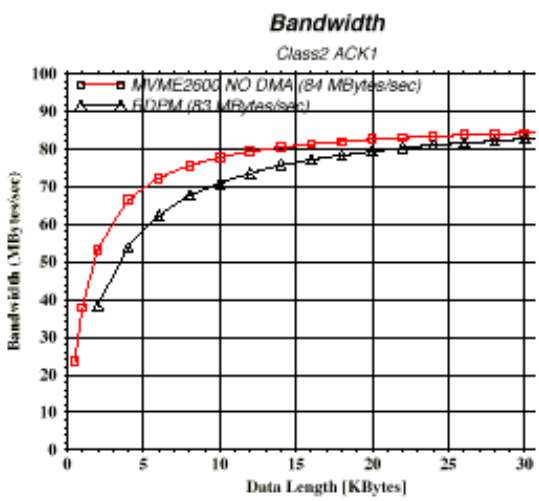
  - received in **Oct. 97** (pre-release)
  - Class 1 at full speed
  - Class 2/3 at full speed
  - frame size = **512 bytes**
  - **NO** compatibility problems with Tachyon 2.0 chip

Switch Overhead









Bottom line:
switch overhead is small
(1 μsec for FC, 0 for ATM)

FC Summary

- **RDPM**

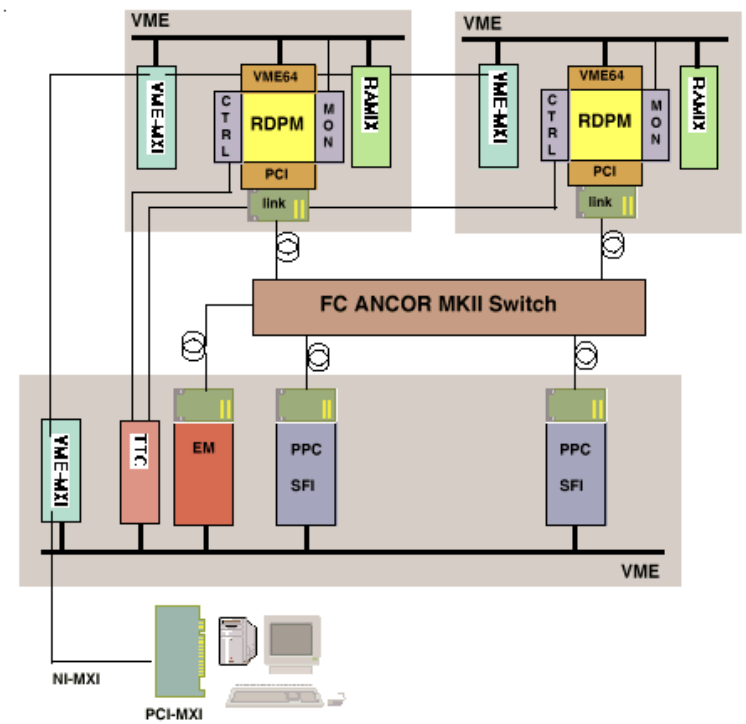
  need for superevents:

  - for 2KBytes blocks (CMS DAQ architecture),
    speed is 42 MBytes/sec in C1
    39 MBytes/sec in C2
  - for large data buffers (32 KBytes),
    speed is 85 MBytes/sec in C1
    84 MBytes/sec in C2

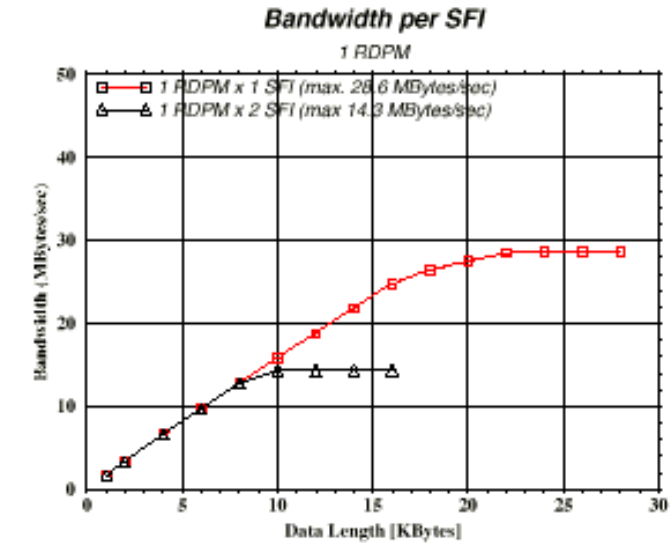
  In C2, the bandwidth curve for RDPMS closely follows the curve of the FC protocol
  (MVME2600 without DMA)

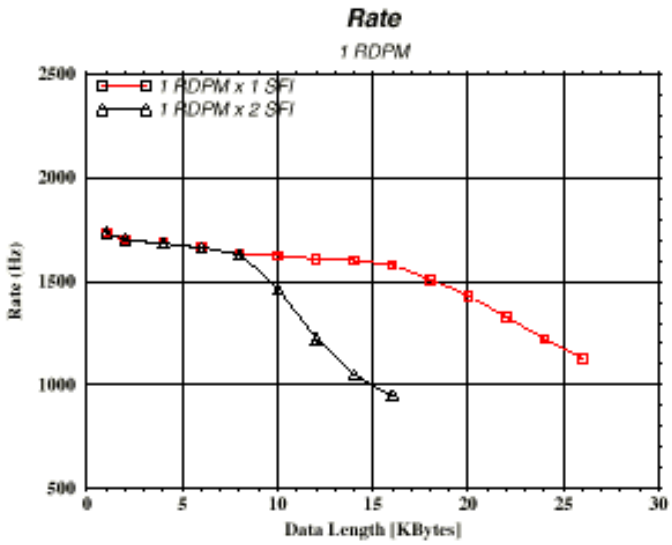
  low latency 22 usecs in C1
  30 usecs in C2

Next: 2x2 DAQ prototype



Linearity/Overheads





Fundamental reason:

must either

tell each source in turn to send data OR

rely on switch to buffer data to the same destination.

(a) Manager doing it: overhead in communication

(b) Switch doing it: buffer limitation (16 kBytes)...

Comparison ATM/FC

- Point-to-Point: FC wins (90 MB/s vs 16 MB/s)

- But no simultaneous sending of events (rate division)

→ in event building pay price 2.5 in throughput

→ 36 MB/s vs 16 MB/s

- But these values depend on event size for FC
  (not for ATM)
  Take scannerSize = 16 kBytes

→ 30 MB/s vs 16 MB/s

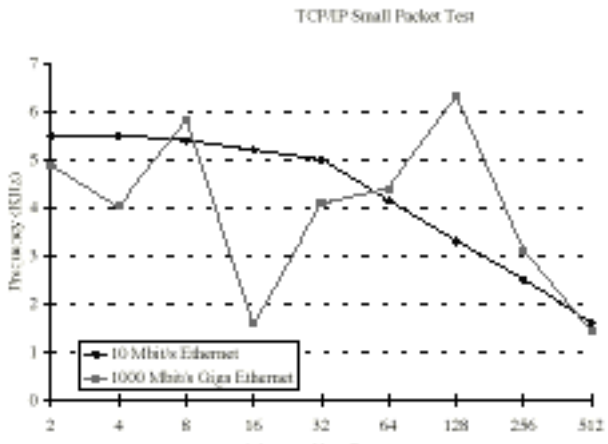
- Add OS overhead (preliminary)

→ 25 MB/s vs 16 MB/s

- Switch availability:
  - ANCOR is non-blocking in Class 1, blocking in Classes 2/3
  - Brocade is non-blocking in all classes but 8x8 for now

Parenthesis: Gbit Ethernet

Note: results preliminary



Erratic behavior not understood

- **Small packets over GE no tuned. Driver problems (??)**
- **TCP/IP transport for level 1 trigger commands (RUI):**

– **Packing 8 LV1 commands in 32 byte packet**

→ **Can have the equivalent of 5x8 = 40 KHz trigger rate. EVMin this case should derandomize the incoming trigger commands, build the ethernet packet and broadcast it to the RUIs. Make sense?**

- **TCP/IP transport for level2 and level 3 (RUO):**

– **In this case probably we can buffer more (and then increase the equivalent rate). Is it true both for level2 and level3? How much we can buffer?**

Is ATM the right technology?

- Basic facts and FAQ (I)
- FAQ (II)
- FAQ (III)
- Decision on ATM (?)
- Conclusion

Basic facts and FAQ (I)

Facts:

- Only 155 Mbit/s (translates to 16 MB/s useful max)
- System works (basically — no intrinsic flaw)
- Market is still wide enough (many manufacturers)
- In point-to-point tests we get 16 MB/s
- Using rate division: linear in N(receivers)
- PCI/PMC choice worked:
  PC farm for output instead of SgI → no change

Question:

- Is this the best Event Builder money can buy?

Answer:

(a) Of course not. We started on this one 4 years ago (!)
(b) The "best" switch comes and goes in ~ 1 year (max)
(c) Another answer: when do we decide? (e.g. now?)
Timescale most important parameter:
Expensive today → cheap in 1 year
Fashionable today → outdated in 3 years

Q: What whould we choose if we were starting today?
A: When do we decide on what to use?

My bias:
Today → ATM
End 98 → Gbit Ethernet
End 99 → Gbit Ethernet or 620 Mb/s ATM
Beg 00 → 2 Gb/s Fibrechannel

FAQ (II)

Q: Is ATM here to stay?
A: Foreseeable future (+3 years) definitely no problem.
ATM was supposed to take over completely, and there was a lot of excitement in the beginning.
The advent of Ethernet has started tipping the scale away from ATM.

Q: Where is the market going?
A: Gigabit Ethernet has largest momentum. Fibrechannel popular but switch market fairly limited

Q: Is FORE reliable? Other suppliers?
A: FORE is big enough. New products continuously out.
Many ATM switch manufacturers. Latest 3COM switch offer is *very* interesting.

Q: Can we upgrade the switch to 620 Mbit/s?
A: Yes, by multiplexing 4 inputs into one module. However, we would do this IFF there was a need for 50 MB/s per scanner.
Then the limitation is VME speed (20 MB/s)

FAQ (III)

Q: How scalable is the system? Is it scalable?
A: Yes, within certain limitations

(a) Switch:
With this type of switch we can have 32x32.
Theoretical maximum is thus 32x16 = 512 MB/s.
Multiply by 1/2 for various ghosts → 256 MB/s.
IFF we believe 300 Hz @ 200 kB → need 60 MB/s.
So, with today's switch, factor 4 safety

(b) Scanner manager/Reflective Memory
This is the safest part of the system
@ 32 nodes, 1 μsec/node → 32 μsec for full circle
Number of messages:
- **1 for L3 box to Manager**
- **1 from manager to Scanners**
- **16 from Scanners to Manager (event done)**
  - **this is equivalent to 8 messages @ 32 μsec**
- **Total of 10 circles → 320 μsec/event**

→ Theoretical maximum = 3000 Hz

(c) VME readout
Probably worst (potential) enemy; Safe bet is 20 MB/s
If we need more → faster CPUs (should go up to 30-40)

Decision on ATM (?)

If we know we don't have to run till 2000, or if we know that the data transport/Level-3 needs will be increased by more than a factor 4, we should consider delaying the decision.

If either of the above is false $\rightarrow$ adopt ATM now.

Conclusion

- Run II System will keep the Run I architecture

- Things which will stay: Reflective Memories, VME

- Things that will change: VME CPUs, Switch, receivers

- We have investigated ATM as a switching technology — and CERN has followed up on Fibrechannel

- Rate Division: the ultimate barrel shifter (most efficient way of putting together an event; only price is increased memory needs on senders)

- ATM stand works. So does the Fibrechannel one. Despite factor 6 in link speed, event building speed is roughly only 50% higher for Fibrechannel (on 2x2). With potential exception of Brocade FC switch, current Fibrechannel switches are blocking beyond 4x4.

- System scales well by up to a factor 4. Beyond that: the brick wall (with FORE, ATM @ 155Mbit/s).

- We can adopt ATM now. However, this system will be running in 2003 also. That's 5 years from now. We should think whether
  - (a) We can wait on decision
  - (b) We plan for (yet another) switch upgrade in the middle of Run II